Georgia Archives

Case Study

Planning for Georgia's Digital Archives

For the Digital Archives, long-term storage must be measured in centuries.

Executive Summary

- **Goals:**
  1. To plan the long-term storage of unknown (but considerable) quantities of data.
  2. To build a flexible solution that does not rely on one vendor or one storage technology.
- **Timetable for planning and implementing the project:** Initial planning and high level design required approximately 3 person weeks. Detailed design, application development and implementation will require an additional 6 person months. Acquisition and installation are expected to require an additional 6 person months.
- **Costs of project:** Costs figures were not available for this case study but planning and design were aimed toward low life-time costs to maximize the return on investment for the state. To accomplish this we are planning a phased approach to the purchase and expansion of the Digital Archives.
- **For more information on this project:** http://www.sos.state.ga.us/archives/who_are_we/rims/digital_History/default.htm

Introduction

The Georgia Archives is responsible for the identification and preservation of the state's recorded history. As state government has relied more and more on electronic processes to conduct business, the Archives must find ways to identify, acquire, and preserve the digital historical records being created. For the last year, the Archives has partnered with the Board of Pardons and Paroles to continue the work begun by the Privacy and Access to Georgia E-Government project of 2003 (case study may be found at http://www.sos.state.ga.us/archives/who_are_we/rims/best_practices_resources/default.htm). Funded by the National Historical Publications and Records Commission, the current project focuses on lifecycle management of electronic records within an agency and the transfer of those permanent records to the Georgia Archives for preservation. The project funded the purchase of servers and storage as the foundation for Georgia's Digital Archives, while project staff worked with the Georgia Technology Authority to identify the functional specifications and requirements of the Digital Archives.

Identifying Storage Solutions for Digital Preservation

Several challenges confronted the project team in planning for the Digital Archives and identifying storage solutions:

- Estimates of storage growth proved very difficult to calculate. A technology survey of file formats used within state government yielded little useful data, and discussions and interviews with agency chief information officers (CIOs) did not produce the expected results. Our own best-guess estimates of intial growth rates for the first five years, based on interviews with agency CIOs and the rate of growth observed with the parole data, are as follows – an initial 1.5 terrabytes, doubling each year (1.5 TB year 1, 3 TB in year 2, 6 TB in year 3, 12 TB in year 4, and 24 TB in year 5). This assumption is based on a slow rate of acquisition of state government data only. Large one-time acquisitions of agency data could alter this estimate as well as the possibility of acquiring local government historical data. Should the opportunity to acquire county historical data occur, these estimates would need to be doubled if not trebled. Most of the storage will be used for databases, image files, and email but some storage will be devoted to web sites. The present estimate of web sites that would be acquired is

approximately 5-10% of the overall number of web sites in state government, or 4-8 sites. Digital audio and video as well as geospatial data are currently excluded from these estimates.

- Evaluating the types of storage needed was another consideration. The system being developed as the Digital Archives consists of three distinct storage components:

    - Preservation storage for the long-term permanent storage of digital objects;
    - Workspace storage for temporary housing of objects during inspection and processing; and,
    - Access storage for temporary housing of records during use.

  The project team elected to focus planning efforts on the preservation and workspace storage components.

- Evaluating program needs beyond the procurement of a storage solution emerged as a related consideration. As the project team was confronted with the realities of storage growth, technology needs, the difficulties of financial sustainability, and existing staffing as compared to future staffing needs, the issue of the program surrounding the storage solution became a central component of the planning process.

The Georgia Archives project team relied heavily on the IT expertise of members from the Board of Pardons and Paroles. The Board's IT infrastructure is far more complex than that of the Archives with over 100 satellite offices throughout the state connected over a wide area network. In addition, the volume of digital records being stored and managed at the Board far exceeded the amounts currently being managed by the Archives. Staff from the Georgia Technology Authority and the Georgia Department of Audits also provided invaluable assistance in identifying hardware and software requirements meeting the Audit Checklist for the Certification of Trusted Digital Repositories, published by the Research Libraries Network in 2005.

Design Considerations

In considering the problem of designing a Digital Archives for the state of Georgia, the project team identified several issues that would inform our design principles:

- The need to manage costs. In evaluating our current budget and the resources we are capable of devoting (at this time) towards the development of a Digital Archives, the project team focused its attention on the construction of the preservation and workspace storage solutions and accepted that compromises in response time may need to be made in order to lower costs.

- The objects to be stored. Our ultimate goal in Georgia is the development of a Digital Records Preservation Coalition – a network of linked digital archives, each specializing in a file format or type of digital object. The project team, therefore, focused its planning around the records of the executive and legislative branches of government. These records are predominantly images, spreadsheets, word processing documents, emails, web sites, and databases (including several large data warehouses). Through the work of the Georgia Technology Authority, file formats in these branches of government are limited to a handful – pdf.s, tiffs,

Exchange Mail, Lotus Notes Mail, Novell GroupWise Mail, Oracle, SQL, Microsoft Office suite, Lotus Notes Office, and Corel Office Suite. By far the dominant applications in government (and becoming even more so each year) are Microsoft and Oracle products.

- The need to design for tolerance to disasters. Hurricane Katrina proved to be a defining moment in the disaster-strewn history of the Southeast. It served, in particular, to point out the absolute necessity of duplicating data outside the immediate area, if not outside the state entirely.

- The need to deploy and manage storage. Part of our strategy for managing costs includes a phased approach to the purchase and installation of the system. The ability to capitalize on decreasing costs of storage and products from multiple vendors requires the development of a flexible storage subsystem capable of accepting components from a diverse vendor group.

- The mechanism of procurement. It is critical that any storage solution selected for the Digital Archives be flexible enough to allow the Archives to manage its costs through the purchase of economical yet robust hardware and software. Therefore, the initial component procurement must not lock the Digital Archives into a relationship with a single vendor.

- The need to plan for long-term storage. To most agencies, long-term storage, at its longest, represents a storage investment in decades. For the Digital Archives, long-term storage must be measured in centuries. The need to plan procurement and manage costs to take advantage of the price decreases and technological advancement is critical to success.

These design principles were further refined into assumptions regarding the functionality and design of the physical storage environment.

- The storage system must be based on a multi-site (at a minimum two sites) design that is not subject to disaster scenarios that are likely in the Southeast (i.e. tornadoes, hurricanes and flooding). These nodes would be on geographically separate sites.

- At least one alternative site must be capable of providing access (even at reduced performance levels) in the event that a disaster prevents the primary site from functioning.

- In order to manage costs, the Digital Archives cannot be locked into a single vendor relationship – no matter how attractive that relationship seems.

- The storage service layer will provide the following functionality:
    1. Allocate unique persistence identifiers to objects;
    2. Bind each of the identifiers permanently to the object;
    3. Guarantee the authenticity and integrity of the object;
    4. Transport each object – reliably – as needed;
    5. Recover from any internal failure; and
    6. Integrate physical storage with external systems.

One final assumption addresses the retrieval rates of the objects in the Digital Archives:

As with the paper Archives holdings, the majority of digital items ingested into the Digital Archives will be retrieved infrequently. Nonetheless, they must be preserved. As the total number of items numbers in the millions, access to even a portion of the holdings will constitute a significant retrieval requirement.

Timetable

As the Georgia Archives moves through its planning process additional design principles and technical requirements will be identified. Below is a timetable for the planning phase of the Digital Archives project:

Objective	Date Completed
Complete NHPRC Grant project by implementing initial storage solution for digital objects	End of May 2006
Define functional specifications for Digital Archives	End of February 2006
Delineate Technology Requirements for Digital Archives	End of April 2006
Release a Request for Information to determine the number of vendors capable of meeting technology requirements	June 2006
Define programmatic needs for the successful implementation of a Digital Archives	June – July 2006
Review/evaluate options and plan for implementation of Digital Archives	July 2006